Parallel approximation of the exponential of Hermitian matrices

Frédéric Hecht, Sidi-Mahmoud Kaber, Lucas Perrin, Alain Plagne, Julien Salomon

PARALLEL APPROXIMATION OF THE EXPONENTIAL OF HERMITIAN MATRICES

FRÉDÉRIC HECHT*, SIDI-MAHMOUD KABER†, LUCAS PERRIN‡, ALAIN PLAGNE§, AND JULIEN SALOMON¶

Abstract. We are interested in the approximation of the exponential function on a real interval by a particular rational function (inverse of a polynomial). Using the partial fraction decomposition of this rational function, we build an algorithm for computing the exponential of a matrix that is adapted to parallel computing. The performances of this parallel algorithm, both in accuracy and in computing time, are quantitatively analyzed.

Key words. Matrix exponential, Parallel computing, Truncation error, Taylor series, Partial fraction decomposition, Padé approximation, MATLAB, Octave, `expm`, Roundoff error.

AMS subject classifications. 15A16, 65F60, 65L99, 65Y05.

1. Introduction. Given a matrix A, the differential equation $u'(t) = Au(t)$ appears in many models, either directly or as an elementary component of more complicated differential systems. To solve with a good accuracy this equation, it is essential to have an algorithm to compute the exponential of a matrix. This algorithm must be efficient, both for the precision and for the time taken to execute it. Such an algorithm is presented in this paper.

Many algorithms exist to compute the exponential of a matrix. We refer the reader to the celebrated review by Moler and Van Loan [10] for a comparison of these methods. None of them is clearly more efficient than the others if we take into account several important criteria such as accuracy, computational time, memory space requirements, complexity, variety of matrices to which the method can be applied, etc. The method we propose in this paper is close to the one defined in [6] or [1], in the sense that on the one hand, it concerns more the action of exponential matrix operator on vectors than the computation of the matrix $\exp(A)$ itself and on the other hand, its objective is to develop calculation algorithms adapted to parallel computing. The algorithm presented in [6] is based on the computation of the exponential of a Heisenberg matrix of dimension m, smaller than the dimension n of the matrix A. This matrix is obtained at the $m-i$-th step of Arnoldi's algorithm. In [1], a scaling and squaring method is used together with a truncated Taylor series approximation to the exponential. Let us also mention the method proposed in [7] which uses a factorization of the matrix A.

In the present work, we present a method based on a simple and old idea which appears to be particularly well adapted to parallel computing. Consider an approximation $\mathcal{E}_n$ of the complex exponential function depending on a parameter $n \in \mathbb{N}$, with $n \geq 1$. This approximation naturally extends to the exponential of diagonal

*Laboratoire Jacques-Louis Lions, Sorbonne Université, CNRS, 75005 Paris and INRIA Paris, ALPINES Project-Team, 75589 Paris Cedex 12, France(frederic.hecht@sorbonne-universite.fr).

†Laboratoire Jacques-Louis Lions, Sorbonne Université, CNRS, 75005 Paris, France (sidi-mahmoud.kaber@sorbonne-universite.fr).

‡INRIA Paris, ANGE Project-Team, 75589 Paris Cedex 12, France and Sorbonne Université, CNRS, Laboratoire Jacques-Louis Lions, 75005 Paris, France (lucas.perrin@inria.fr).

§Centre de Mathématiques Laurent Schwartz, École polytechnique, F-91128 Palaiseau, France. (alain.plagne@polytechnique.edu).

¶INRIA Paris, ANGE Project-Team, 75589 Paris Cedex 12, France and Laboratoire Jacques-Louis Lions, Sorbonne Université, CNRS, 75005 Paris, France (julien.salomon@inria.fr).

matrices by setting $\mathcal{E}_n(\mathrm{diag}(d_i)_i) = (\mathrm{diag}(\mathcal{E}_n(d_i))_i$, thus to any diagonalizable matrix $A = P^{-1}DP$ by $\mathcal{E}_n(A) = P\mathcal{E}_n(D)P^{-1}$. The approximation error for a diagonalizable matrix $A = PDP^{-1}$ can then be estimated as follows

$$\|\exp(A) - \mathcal{E}_n(A)\|_2 \leq \kappa_2(P)\,\|\exp(D) - \mathcal{E}_n(D)\|_2,$$

where $\kappa_2(P) = \|P\|_2\,\|P^{-1}\|_2$ is the condition number of P in the matrix norm associated with the usual Euclidean vector norm $\|\cdot\|_2$. Let Λ be a domain of the complex plane that contains the spectrum of A. It follows that if there exists a sequence $(\varepsilon_n)_n > 0$ converging rapidly to 0 such that $\max_{z\in\Lambda} |\mathcal{E}_n(z) - \exp(z)| \leq \varepsilon_n$, then, for any diagonalizable matrix whose spectrum is included in Λ

$$\|\exp(A) - \mathcal{E}_n(A)\|_2 \leq \varepsilon_n\,\kappa_2(P)\,. \tag{1.1}$$

The approximation of $\exp(A)$ is then reduced to the approximation of the exponential on the complex plane. If we further assume A to be Hermitian (or, more generally, normal), then P is a unitary matrix and $\kappa_2(P) = 1$. This is the case for Hermitian matrices, and more generally for normal matrices. Note, however, that for an arbitrary matrix, the term $\kappa_2(P)$ may be too large and seriously degrade the estimate (1.1). As we are interested in matrices coming from the Laplacian discretization, we will consider in this work only Hermitian matrices.

In our approach, the approximation $\mathcal{E}_n$ is based on the the partial fraction decomposition of $1/\exp_n(-x)$, where $\exp_n$ denotes the truncated Taylor series of order n associated with the exponential. All terms in this decomposition are independent hence their computation can achieved efficiently in parallel.

The article is organized as follows. Section 2 is devoted to the approximation of the scalar exponential function. This approximation, denoted by $\mathcal{R}_n(z)$ is in our approach a rational function whose poles are all simple. In Section 3, we present the approximation of the exponential of a matrix. In practise, the partial fraction decomposition of $\mathcal{R}_n(z)$ at the basis of our approach raises some specific numerical issues related to floating-point arithmetic ; these are discussed in Section 4. The efficiency of our method is demonstrated on some examples in Section 5. Some concluding remarks are given in Section 6.

2. The scalar case. For $n \in \mathbb{N}^*$, let us define $\exp_n(z) := \sum_{k=0}^{n} \frac{1}{k!} z^k$, i.e., the exponential Taylor series truncated at order n. For any complex z, we have $\lim_{z\to+\infty} \exp_n(z) = \exp(z)$. It is readily seen that for all $x \in \mathbb{R}$ and even values of n, $\exp_n(x)$ is positive. Since $\exp_n' = \exp_{n-1}$, it follows that $\exp_n$ is strictly increasing for n odd and strictly convex for n even.

2.1. Roots of the truncated exponential series. We denote by $(\theta_k^{(n)})_{k=1,\cdots,n}$ the roots of the polynomial $\exp_n$. If n is even, the roots are pairs of conjugate complex numbers and none of them is a real number. If n is odd, there is one and only one real root of $\exp_n$ and the others are pairwise conjugate. Some roots of $\exp_n$ are represented on the figure 1 (left panel). We see that the norm of the roots increases with n, which intuitively follows from the fact that the exponential function has no roots on the whole complex plane. However, this growth is moderate since (see [15], for example)

$$1 \leq |\theta_k^{(n)}| \leq n. \tag{2.1}$$

G. Szegőe has shown in [13] that the *normalized roots*, *i.e*, the roots of $\exp_n(nz)$,

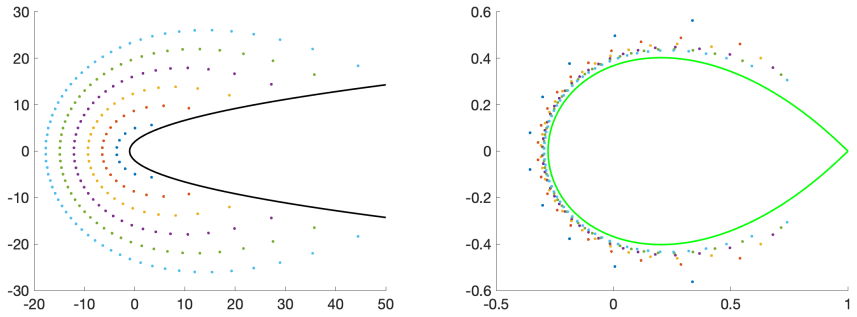

Figure 1: Left: the roots of $z \longmapsto \exp_n(z)$, $n = 10, 20, 30, 40, 50, 60$. The parabola $y^2 = 4(x+1)$ delimits an area containing no roots. Right: the Szegőe curve.

approach, when $n \to \infty$, the so-called Szegőe curve, defined by

$$\{z \in \mathbb{C}, \quad |ze^{1-z}| = 1, \quad |z| \leq 1\}.$$

Some normalized roots and the Szegőe curve are presented in Figure 1 (right panel). In view of (1.1), it is interesting to determine parts of the complex plane which do not contain any roots. An example is given by the interior of the parabola of equation $y^2 = 4(x+1)$, which thus includes the positive real half-axis. This surprising result has been obtained by Saff and Varga in [11], see Figure 1 (left).

2.2. Approximation of the exponential. Considering the identity $\exp(z) = \frac{1}{\exp(-z)}$, we propose the following approximation of the exponential function defined for any complex number z by

$$\exp(z) \simeq \mathcal{R}_n(z) := \frac{1}{\exp_n(-z)}.$$

Note that $\mathcal{R}_n(0) = 1$ and that $\mathcal{R}_n$ has no real root if n is even, which we will always assume in the rest of this paper.

This approach opens the way to a good approximation of the exponential on the half axis $(-\infty, 0]$. We present on Figure 2 a graph of the exponential function exponential function and its polynomial approximation $\exp_n$ and rational approximation $\mathcal{R}_n$, on the interval $[-5, 0]$. We observe that for $n = 10$, the rational approximation is clearly more accurate. For $n = 20$, the two approximations seem to fit well with the exponential function.

Remark 2.1. *Given $n, m \in \mathbb{N}$, the Padé approximant [2] of index (m, n) of the exponential function is explicitly known ; it is the rational function with numerator $P_{m,n}$ and denominator $Q_{m,n}$:*

$$P_{m,n}(x) = \sum_{k=0}^{m} \frac{(m+n-k)!m!}{(m+n)!(m-k)!k!} x^k, \quad Q_{m,n}(x) = \sum_{k=0}^{n} \frac{(m+n-k)!n!}{(m+n)!(n-k)!k!} (-x)^k.$$

Function $\mathcal{R}_n(z)$ is therefore the Padé approximant of index $(0, n)$ of the exponential function. Its Taylor expansion at the origin coincides with this function up to order n. More precisely, we have

$$\mathcal{R}_n^{(j)}(0) = 1, \qquad j \in \mathbb{N}, \ 0 \leq j \leq n, \tag{2.2}$$

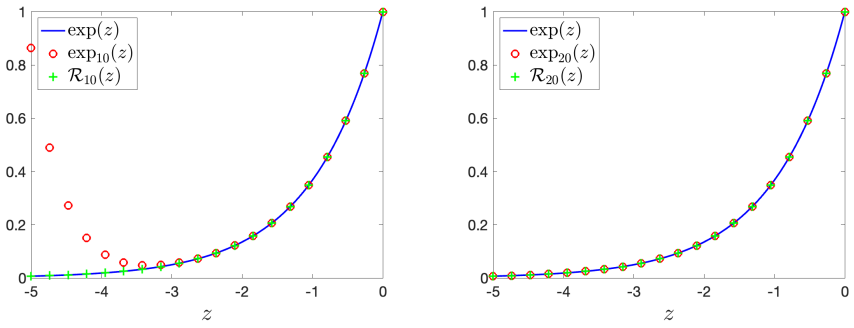

Figure 2: The exponential function and its polynomial and rational approximations rational $\exp_n$ and $\mathcal{R}_n$ on $[-5, 0]$, for $n = 10$ (left) and $n = 20$ (right)

and, in the neighborhood of the origin,

$$\mathcal{R}_n(z) = \exp(z) + \mathcal{O}(z^{n+1}).$$

We can refine this result a little. Let us decompose $\mathcal{R}_n(z)$ as

$$\mathcal{R}_n(z) = \exp_n(z) + \sum_{k=n+1}^{+\infty} \frac{\lambda_{n,k}}{k!} z^k.$$

A simple calculation shows that (recall that n is supposed to be even) $\lambda_{n,n+1} = 0$ and $\lambda_{n,n+2} = -2(n+1)$. In other words, in $z = 0$, the derivative of $\mathcal{R}_n(z)$ of order greater than n are not at all close to the derivatives of the exponential function.

The partial fraction decomposition of $\mathcal{R}_n$ is the basis of our numerical method to compute the exponential of a matrix.

PROPOSITION 2.2. *We have, for all* $z \in \mathbb{C}$

$$\mathcal{R}_n(z) = \sum_{k=1}^{n} \frac{a_k^{(n)}}{z + \theta_k^{(n)}}, \tag{2.3}$$

where $\theta_k^{(n)}$ *are the roots of* $\exp_n$ *and*

$$a_k^{(n)} = -\frac{n!}{\prod_{j \neq k} (\theta_k^{(n)} - \theta_j^{(n)})}. \tag{2.4}$$

One should not be alarmed in the calculation of the coefficients $a_k^{(n)}$ by the relation (2.4) whose denominator is a product of the differences $\theta_k^{(n)} - \theta_j^{(n)}$ because the difference between two roots of $\exp_n$ is uniformly lower bounded with respect to n (see [15, Theorem 4])

$$\inf_{n \geq 2} \min_{j \neq k} |\theta_j^{(n)} - \theta_k^{(n)}| \geq \gamma := 0.29044\cdots \tag{2.5}$$

thus avoiding to divide by too small numbers in (2.4). Moreover, other expressions can be used to compute the coefficients $a_k^{(n)}$, e.g.,

$$a_k^{(n)} = \frac{-1}{\exp_n'(\theta_k^{(n)})} = \frac{n!}{(\theta_k^{(n)})^n}. \tag{2.6}$$

2.3. Convergence and error estimate. The rational approximation of a real or complex function is a well-documented problem: existence of a better approximation, uniqueness, computation, etc. See, for example [9]. Two problems arise.

1. Determine a part Λ of the complex plane where the approximation $\mathcal{R}_n(z) \simeq \exp(z)$ converges:
$$\lim_{n\to+\infty} |\mathcal{R}_n(z) - \exp(z)| = 0, \quad (\forall z \in \Lambda).$$
Obviously, one needs to exclude neighborhoods of the poles of $\mathcal{R}_n$
2. Determine parts of the complex plane where the approximation is accurate, *i.e.* where the error $|\mathcal{R}_n(z) - \exp(z)|$ rapidly goes to 0 when n goes to infinity, e.g., linearly.

Let us denote $\mathbb{P}_{m,n}(\Lambda)$ the set of rational functions defined on Λ whose numerator and denominator are of degree m (at most) and n (at most) respectively and define

$$E_{m,n}(\exp, \Lambda) = \min_{r\in\mathbb{P}_{m,n}(\Lambda)} \max_{z\in\Lambda} |\exp(z) - r(z)|,$$

the error of best uniform approximation on Λ of the exponential function by $\mathbb{P}_{m,n}(\Lambda)$.

2.3.1. Convergence on a bounded domain. In [3], one obtains

$$E_{m,n}(\exp, [-1,1]) \underset{n+m\to\infty}{\simeq} \frac{n!m!}{2^{n+m}(n+m)!(n+m+1)!}.$$

In this work we are only interested in the approximation in $\mathbb{P}_{0,n}$ for which we have

$$E_{0,n}(\exp, [-1,1]) \underset{n\to\infty}{\simeq} \frac{1}{2^n(n+1)!}.$$

Such strong decrease can however only be obtained on bounded intervals of $\mathbb{R}$. This is not the case in the applications which have motivated our study. The behavior of the approximation error on a disk has been also analysed [14]:

$$E_{0,n}(\exp, B(0,\varrho)) \underset{n\to\infty}{\simeq} \frac{1}{(n+1)!}\varrho^{n+1},$$

where $B(0,\varrho)$ denotes the ball of the complex plane centred in 0 and of radius ϱ.

2.3.2. Convergence on $]-\infty, 0]$. This case is treated in the pioneering work of [4] where $\exp(-x)$ is approached for positive values of x. The authors show that the best approximation error $E_{0,n}(\exp,]-\infty, 0])$ decays linearly and exhibit a particular function, which just happens to be our rational approximation $\mathcal{R}_n$.

PROPOSITION 2.3. *([4, Lemma 1]) We have for any real $x \leq 0$*

$$|\mathcal{R}_n(x) - \exp(x)| \leq \frac{1}{2^n}. \tag{2.7}$$

The convergence of $\mathcal{R}_n$ is largely sufficient in practical applications. For the sake of completeness, let us note that the optimal linear decrease is given by Schönhage in [12] who showed that $\lim_{n\to+\infty} E_{0,n}^{1/n}(\exp,]-\infty, 0]) = 1/3$.

The convergence of $\mathcal{R}_n(x)$ to $\exp(x)$ is therefore linear on the half-line of the negative half-line as we can see on the curves in Figure 3. Figure 4 shows iso-curves of the norm of the error for $n = 32$ as well as points $-\theta_k^{(n)}$. The iso-curves of the

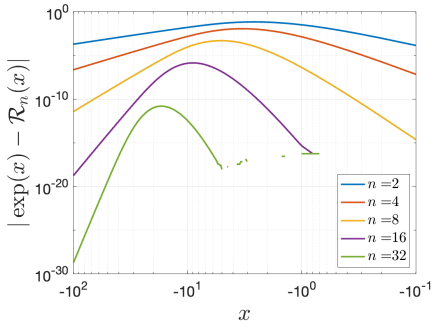

Figure 3: The error $|\mathcal{R}_n(x) - \exp(x)|$ for $n \in \{4, 8, 16, 32\}$.

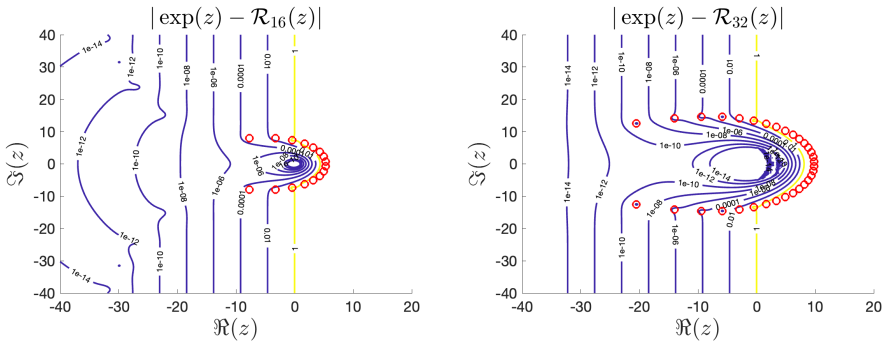

Figure 4: *Norm of error* $\mathcal{R}_n(z) - \exp(z)$ *and poles of* $\mathcal{R}_n$ *for* $n = 16$ *(left) and* $n = 32$ *(right)*

norm of error for $n = 32$ and the points $-\theta_k^{(n)}$. The iso-curves of the norm of error for $n = 32$ are the 10^{-k} for $k = 0, \cdots, 14$. One observes there the rapid decay of the approximation along the real half-axis. We also observe a remarkable decay in the whole left half-plane.

Before going further, let us state a technical result.

LEMMA 2.4. *The function* $f_n(x) := \exp_n(x)\exp(-x)$ *satisfies*

$$f_n(n+1) < \frac{1}{2} \tag{2.8}$$

$$f_n^2(x) \leq \frac{n+1}{n} f_{n-1}(x) f_{n+1}(x). \tag{2.9}$$

Proof. We first show that

$$\exp_n(n+1) = \sum_{k=0}^{n} \frac{(n+1)^k}{k!} < \exp(n+1) - \exp_n(n+1) = \sum_{k=n+1}^{+\infty} \frac{n^k}{k!}, \tag{2.10}$$

which directly leads to (2.8). To get (2.10), we compare the terms $k = n - j$ and $k = n + j - 1$ of the respective sums. Precisely, we have

$$\frac{n^{n-j}}{(n-j)!} \leq \frac{n^{n+j-1}}{(n+j-1)!}.$$

The proof is by induction on j. For $j=1$, we actually have equality. Assuming the property is true at rank j, we have

$$\begin{aligned}\frac{n^{n-(j+1)}}{(n-(j+1))!} =& \frac{n^{n-j}}{(n-j)!}\frac{n-j}{n} < \frac{n^{n+j-1}}{(n+j-1)!}\frac{n-j}{n} = \frac{n^{n+j}}{(n+j)!}(1-(\frac{j}{n})^2)\\ <& \frac{n^{n+j}}{(n+j)!},\end{aligned}$$

hence the result. Inequality (2.9) simply follows from the Cauchy-Schwarz inequality applied to

$$\int_x^{+\infty} \exp(-t)t^n dt = \int_x^{+\infty} (\exp(-t/2)t^{\frac{n-1}{2}})\,(\exp(-t/2)t^{\frac{n+1}{2}})\,dt,$$

since $f_n(x) = \int_x^{+\infty} \exp(-t)\frac{t^n}{n!}dt$. ∎

We summarize some properties of the function $err_n : x \in]-\infty, 0] \longmapsto \mathcal{R}_n(x) - \exp(x) > 0$ in the following proposition.

PROPOSITION 2.5. *For $n \geq 1$, the function err_n reaches its maximum at a single point $\xi_n < 0$, is increasing on $]-\infty, \xi_n]$ and decreasing on $[\xi_n, 0]$. Moreover, we have*

$$\frac{n}{2} \leq -\xi_n \leq n+2. \tag{2.11}$$

Proof. To simplify the notation, we introduce $y = -x$ and study the error $err_n(y) = \mathcal{R}_n(-y) - \exp(-y)$ on the half-axis $(0, +\infty[$, in which we have excluded $y = 0$ where the error cancels. Since

$$\begin{aligned}err_n'(y) =& \exp(2y)\frac{(\exp_n(y)\exp(-y))^2 - \exp_{n-1}(y)\exp(-y)}{\exp_n(y)^2} \\ =& y^n \frac{\exp(-y)}{\exp_n(y)^2}\left(\frac{1}{n!}\exp_n(y) - \exp_{n-1}(y)\frac{\exp(y)-\exp_n(y)}{y^n}\right),\end{aligned} \tag{2.12}$$

the variations on err_n are determined by the sign of

$$g_n(y) := \frac{1}{n!}\exp_n(y) - \exp_{n-1}(y)\frac{\exp(y)-\exp_n(y)}{y^n}. \tag{2.13}$$

In this formula, the last term as well as all its derivatives is positive on $I =]0, +\infty[$ so that $g_n^{(n+1)}(y) < 0$ on this interval. To prove that g_n has an unique zero ξ_n in I, we shall show that for some $a_1 > 0$, g_n' is strictly positive on in interval $]0, a_1[$ and strictly negative on $]a_1, +\infty[$. That guarantees the result since $g_n(0) > 0$ and $\lim_{y\to+\infty} g_n(y) = -\infty$. The values of $g_n^{(k)}$ $(k = 1, \cdots, n)$ on both sides of interval I are of importance in the analysis. Note first that $\lim_{y\to+\infty} g_n^{(k)}(y) = -\infty$ for all $k \in \mathbb{N}$ and that the sequence $u_k := g_n^{(k)}(0) = \frac{1}{n!} - \sum_{\ell=1}^k \frac{k!}{(k-\ell)!(n+\ell)!}$ is decreasing. Indeed, for $k = 1, \cdots, n-1$ we have

$$\begin{aligned}u_{k+1} - u_k =& -\frac{(k+1)!}{(n+k+1)!} + \sum_{\ell=1}^{k}\frac{k!}{(k-\ell)!(n+\ell)!} - \frac{(k+1)!}{(k+1-\ell)!(n+\ell)!} \\ =& -\frac{(k+1)!n!}{(n+k+1)!} - k!\sum_{\ell=1}^{k}\frac{1}{(k-\ell)!(n+\ell)!}\frac{\ell}{k+1-\ell} < 0.\end{aligned}$$

Since the first term in the right hand side of (2.13) is a polynomial of order n, we have $g_n^{(n+1)}(y) < 0$ on I. Hence $g_n^{(n)}$ is strictly decreasing on this interval. If $u_n > 0$, then for some $a_n > 0$, $g_n^{(n)} > 0$ on some interval $[0, a_n[$ and $g_n^{(n)} < 0$ on $]a_n, +\infty[$. Hence function $g_n^{(n-1)}$ is strictly increasing on the first interval and decreasing on the second one. It follows that there exists a unique $a_{n-1} > a_n$ where this function vanishes. This property clearly spreads to g_n. Otherwise, $u_n \leq 0$, and the previous reasoning can be applied to the largest n' such that $u_{n'} > 0$.

To prove (2.11), we first show that $err'_n(n/2) < 0$, i.e., $g_n(n/2) > 0$. Set $y = n\theta$, with $0 < \theta < 1$. We have

$$\begin{aligned}\frac{\exp(y) - \exp_n(y)}{y^n} =& \frac{1}{(n\theta)^n} \sum_{\ell=n+1}^{+\infty} \frac{\theta^\ell}{\ell!/n^\ell} \\ \leq& \frac{1}{(n\theta)^n} \frac{n^{n+1}}{(n+1)!} \sum_{\ell=n+1}^{+\infty} \theta^\ell = \frac{n}{(n+1)!} \frac{\theta}{1-\theta}.\end{aligned}$$

As a consequence, we get

$$g_n(n\theta) \geq \frac{(n\theta)^n}{(n!)^2} + \frac{\exp_{n-1}(n\theta)}{n!} \left(1 - \frac{n}{n+1} \frac{\theta}{1-\theta}\right),$$

which is positive when $\theta = 1/2$.

We then prove that $err'_n(n+2) < 0$. Rewriting (2.12) with the notation of Lemma 2.4, we get $err'_n(y) = \exp(2y) \frac{(f_n(y))^2 - f_{n-1}(y)}{\exp_n(y)^2}$. The task is now to find the sign of $(f_n(n+2))^2 - f_{n-1}(n+2)$. Because of Lemma 2.4, we have

$$\begin{aligned}(f_n(n+2))^2 - f_{n-1}(n+2) \leq& \left(\frac{n+1}{n} f_{n+1}(n+2) - 1\right) f_{n-1}(n+2) \\ \leq& \left(\frac{n+1}{2n} - 1\right) f_{n-1}(n+2),\end{aligned}$$

where the former inequality follows from (2.9) and the latter from (2.8). This shows that $(f_n(n+2))^2 - f_{n-1}(n+2) < 0$. Hence $err'_n(n+2) < 0$. ∎

3. Approximation of the exponential of Hermitian matrices. Let A be a square matrix of $\mathcal{M}_d(\mathbb{C})$. Given $n > 1$, we suppose that all matrices $A - \theta_k^{(n)} I$ are invertible, i.e., their spectrum does not contain any root of any $\exp_n$. This is the case if the matrix A is Hermitian (recall that n is supposed to be even). And the same is true for any matrix provided that n is large enough. We propose the following approximation of the exponential of A

$$\exp(A) \simeq \mathcal{R}_n(A) := \sum_{k=1}^{n} a_k^{(n)} (A + \theta_k^{(n)} I)^{-1}, \tag{3.1}$$

where I denotes the identity matrix.

REMARK 3.1. *Note that $\mathcal{R}_n(0) = I$ and that if the matrix $D \in \mathcal{M}_d(\mathbb{C})$ is diagonal, so is matrix $\mathcal{R}_n(D)$ with $(\mathcal{R}_n(D))_{i,i} = \mathcal{R}_n(D_{i,i})$. On the other hand, for any invertible matrix $P \in \mathcal{M}_d(\mathbb{C})$, we have*

$$\mathcal{R}_n(PAP^{-1}) = P\mathcal{R}_n(A)P^{-1}.$$

From now on, we restrict our attention to negative Hermitian matrices. In view of Proposition 2.5, we can state a specific estimate in this case.

THEOREM 3.2. *Assume that* $Spec(A) \subset \mathbb{R}^-$ *and* $n > 2\varrho(A)$ *with*

$$\varrho(A) := \max_{\lambda \in Spec(A)} |\lambda|,$$

then

$$\| \exp(A) - \mathcal{R}_n(A) \|_2 \leq \varepsilon,$$

where $\varepsilon = \mathcal{R}_n(\varrho(A)) - \exp(\varrho(A))$.

REMARK 3.3 (Shifting method for nonnegative Hermitian matrices). *Since the spectrum of a real-valued matrix can be localized everywhere in the complex plane, we cannot guarantee that* (2.7) *holds in the general case. This problem can be solved by a shifting method in the case of Hermitian matrices. Let* A *be an Hermitian matrix and* $c \in \mathbb{R}$ *a bound of its spectrum,* $c \geq \alpha(A) := \max_i \lambda_i$. *Since* $Spec(A - cI) \subset \mathbb{R}^-$, *we can consider the approximation*

$$\exp(A) = e^c \exp(A - cI) \simeq e^c \, \mathcal{R}_n(A - cI).$$

But the term e^c *can be very large so that the approximation is only reasonable for moderate values of* c. *However the relative error*

$$\frac{\| \exp(A) - e^c \, \mathcal{R}_n(A - cI) \|_2}{\| \exp(A) \|_2} \leq \frac{e^c}{\| \exp(A) \|_2} \| \exp(A - cI) - \mathcal{R}_n(A - cI) \|_2$$

is under control since for Hermitian matrices, $\| \exp(A) \|_2 = e^{\alpha(A)}$.

Computation of $\exp(A)v$. Given $v \in \mathbb{C}^d$, $y = \exp(A)v$ is computed by

$$y \simeq \mathcal{R}_n(A)v = \sum_{k=1}^{n} a_k^{(n)} y_k^{(n)}, \tag{3.2}$$

with $y_k^{(n)}$ the solution to the linear system $(A + \theta_k^{(n)} I) y_k^{(n)} = v$. Each $y_k^{(n)}$ could be computed separately from the others leading to significant savings in computational time as illustrated by our numerical tests, see Section 5.

4. Floating-point arithmetic and numerical implementation. In this section, we examine the efficiency of the approximation

$$\exp(x) \simeq \sum_{k=1}^{n} \frac{a_k^{(n)}}{x + \theta_k^{(n)}},$$

where x is assume to be a real number. We decompose the error according to

$$\underbrace{\exp(x) - \sum_{k=1}^{n} \frac{a_k^{(n)}}{x + \theta_k^{(n)}}}_{e_1(x)} = \underbrace{\left(\exp(x) - \frac{1}{\exp_n(-x)} \right)}_{e_2(x)} + \underbrace{\left(\frac{1}{\exp_n(-x)} - \sum_{k=1}^{n} \frac{a_k^{(n)}}{x + \theta_k^{(n)}} \right)}_{e_3(x)}.$$

The latter cancels in exact arithmetic. However, working for example in a finite precision of 16 significant digits, we see on Figure 5 (left) that in practise e_1 decreases until approximately $n = 34$ and then increases. This behavior makes our approximation

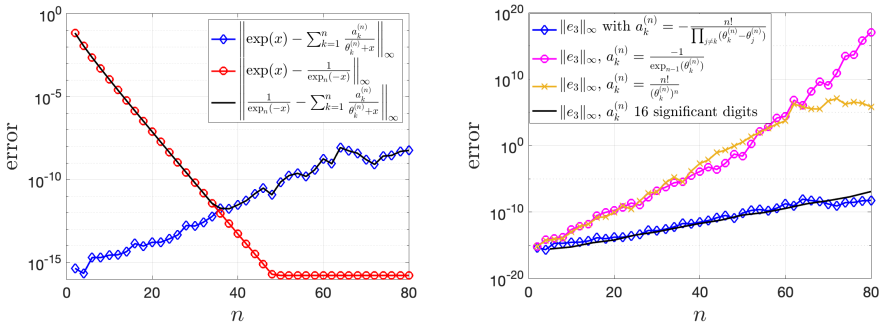

Figure 5: Left: uniform norm of e_1, e_2, and e_3 over $[-100, 0]$ as a function of n. Right: norm of e_3 using various definitions of $a_k^{(n)}$, the black curve computation is done with $a_k^{(n)}$ computed with 32 significant digits and truncated to 16 significant digits.

not accurate and our approach uninteresting in practise for large values of n. These two behaviours can be explained by e_2 and e_3, respectively. The former decreases as predicted by Proposition 2.3. The latter increases with respect to n. The increase in e_3 is related to the partial fraction decomposition in floating-point arithmetic. In what follows, we use Matlab [8] and Octave [5], with double precision. The accuracy actually deteriorates for larger values of n. Indeed, the three equivalent definitions of the coefficients $a_k^{(n)}$ given by (2.4) and (2.6) lead in practise to different numerical results. The uniform norm of e_3 obtained with each of these definitions is shown on Figure 5 (right). The formula given in (2.4) gives the most precise results, which actually very similar to the one obtained by keeping the exact 16 first digits. Hence, we use (2.4) in the sequel.

In order to understand the influence of the floating-point arithmetic, we present a numerical bound whose graph is plotted in Figure 6. This bound guarantees a certain precision for a given n when working with a floating-point arithmetic of D significant digits. We see in that example that with 16 significant digits, we can choose $n = 30$ to guarantee an error of order 10^{-8} and get actually order 10^{-10}.

PROPOSITION 4.1. *Denote by $\widetilde{a}_k^{(n)}$ and $\widetilde{\theta}_k^{(n)}$, the D-significant digits approximations of $a_k^{(n)}$ and $\theta_k^{(n)}$, and assume that*

$$\gamma > n10^{(1-D)} \tag{4.1}$$

with γ defined in (2.5). We have the following upper bound, for $x \in \mathbb{R}$:

$$\left| \frac{1}{\exp_n(-x)} - \sum_{k=1}^{n} \frac{\widetilde{a}_k^{(n)}}{\widetilde{\theta}_k^{(n)} + x} \right| \leq (C_1(n, D) + C_2(n, D)) \sum_{k=1}^{n} \left| \widetilde{a}_k^{(n)} \right|, \tag{4.2}$$

where

$$C_1(n, D) := \frac{2 \cdot 10^{(1-D)}}{\gamma \cdot \left(1 - 10^{(1-D)}\right)}, \quad C_2(n, D) := \frac{4n \cdot 10^{(1-D)}}{\gamma |\gamma - n \cdot 10^{(1-D)}|}.$$

Note that (4.1) is not restrictive: it holds for example in the case of 16 significant digits even if $n \approx 10^{10}$.

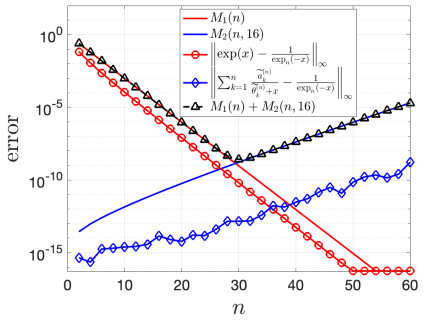

Figure 6: Uniform norm of e_2 and e_3 over $[-100, 0]$, in 16 significant digits, and upper bounds : $M_1(n) = 1/2^n$ and $M_2(n, D = 16)$ given by (4.2).

Proof. Since we are dealing with numerical approximations based on D significant digits, we consider the first D digits of $\widetilde{a}_k^{(n)}$ and $\widetilde{\theta}_k^{(n)}$ to be exact. Then, for any complex number z and its approximation $\widetilde{z}$ we have:

$$\widetilde{z} = z(1+\varepsilon_z), \quad |\varepsilon_z| \in \left[10^{-D}, 10^{(1-D)}\right] \tag{4.3}$$

Writing $r_k^{(n)}(x) = \frac{1}{\theta_k^{(n)}+x}$ and $\widetilde{r}_k^{(n)}(x) = \frac{1}{\widetilde{\theta}_k^{(n)}+x}$, we see that finding an upper bound for the left side of (4.2) amounts to finding an upper bound for:

$$\begin{aligned}
&\sum_{k=1}^{n} r_k^{(n)}(x)\left(\widetilde{a}_k^{(n)} - a_k^{(n)}\right) + \widetilde{a}_k^{(n)}\left(\widetilde{r}_k^{(n)}(x) - r_k^{(n)}(x)\right) \\
&= \sum_{k=1}^{n} r_k^{(n)}(x) a_k^{(n)} \left(\frac{\varepsilon_{a_k^{(n)}}}{1+\varepsilon_{a_k^{(n)}}}\right) + \sum_{k=1}^{n} \widetilde{a}_k^{(n)}\left(\widetilde{r}_k^{(n)}(x) - r_k^{(n)}(x)\right),
\end{aligned}$$

where the equality follows from (4.3). Combining (2.1) with (4.3), we get $|\widetilde{\theta}_k^{(n)} - \theta_k^{(n)}| = |\varepsilon_{\theta_k^{(n)}}\theta_k^{(n)}| \leq n \cdot 10^{(1-D)}$. Combining (2.5) with the fact that for n even, $\theta_k^{(n)}$ are strictly not real, we obtain that $\left|\theta_k^{(n)} + x\right| \geq \left|\mathcal{I}m(\theta_k^{(n)})\right| \geq \frac{\gamma}{2}$ when $x \in \mathbb{R}$. As a consequence $\left|r_k^{(n)}(x)\right| \leq \frac{2}{\gamma}$ for all $x \in \mathbb{R}$. In the same manner, we can see that

$$\left|\widetilde{\theta}_k^{(n)} + x\right| \geq \left|\mathcal{I}m(\widetilde{\theta}_k^{(n)})\right| \geq \left||\mathcal{I}m(\theta_k^{(n)})| - |\mathcal{I}m(\varepsilon_{\theta_k^{(n)}}\theta_k^{(n)})|\right| \geq \frac{\left|\gamma - n10^{(1-D)}\right|}{2}$$

which follows from (4.1). Consequently, $\left|\widetilde{r}_k^{(n)}(x)\right| \leq \frac{2}{\left|\gamma - n10^{(1-D)}\right|}$ for all $x \in \mathbb{R}$.

Finally, we have $|\widetilde{r}_k^{(n)}(x) - r_k^{(n)}(x)| = |\widetilde{r}_k^{(n)}(x)||r_k^{(n)}(x)||\widetilde{\theta}_k^{(n)} - \theta_k^{(n)}|$ so that $|\widetilde{r}_k^{(n)}(x) - r_k^{(n)}(x)| \leq \frac{4n \cdot 10^{(1-D)}}{\gamma\left|\gamma - n10^{(1-D)}\right|}$. Combining all theses inequalities with $|\varepsilon_{a_k^{(n)}}| \leq 10^{(1-D)}$, and $\left|\frac{\varepsilon_{a_k^{(n)}}}{1+\varepsilon_{a_k^{(n)}}}\right| \leq \frac{10^{(1-D)}}{1-10^{(1-D)}}$, we get (4.2). ∎

The graphs of $e_1(x)$ for $x = -10$ obtained with various number of significant digits is given in Figure 7. We see that the larger the number of significant digits, the

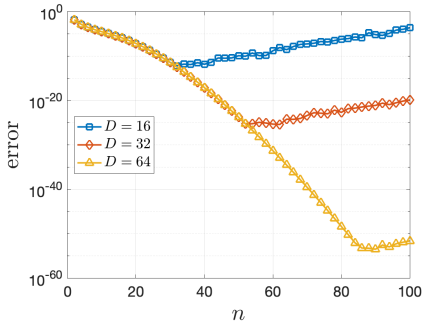

Figure 7: Error (e_1) on the approximation of $\exp(x)$ vs n ($x = -10$) computed with Maple, with 16, 32 and 64 significant digits, respectively.

later e_1 starts increasing. It follows that floating-point arithmetic precision must be adapted to n which is in practise the number of processor used in the computation.

REMARK 4.2. *If n is even and $x \in \mathbb{R}$, we can compute twice as fast $\mathcal{R}_n(x)$. Indeed, the complex numbers $\theta_k^{(n)}$ are in this case a set of conjugate pairs as well as $a_k^{(n)}$, and $\frac{1}{\theta_k^{(n)}+x}$. Assuming that the labelling is such that $\theta_{2\ell+1}^{(n)} = \overline{\theta_{2\ell}^{(n)}}$, we get*

$$\sum_{k=1}^{n} \frac{a_k^{(n)}}{x+\theta_k^{(n)}} = \sum_{\ell=1}^{n/2} 2\mathcal{R}e\left(\frac{a_{2\ell}^{(n)}}{x+\theta_{2\ell}^{(n)}}\right).$$

It follows that the number of computations can be divided by two. The same holds for the computation of $\mathcal{R}_n(A)$ when $Spec(A)$ is real. However, the condition number of transition matrix the may significantly increases in this case.

5. Numerical efficiency. As a first example, we consider a symmetric $d \times d$ real matrix with spectrum randomly chosen in $[-50, 0]$. The relative error $\frac{\|\exp(A)-\mathcal{R}_n(A)\|_2}{\|\exp(A)\|_2}$ as a function of the dimension d is represented in Figure 8 (left). The results are smoothed by using the mean of the error for various random spectra. We use the approximation (3.1) where the inverse matrix is computed with the functions `inv` of Matlab and Octave. We note that the error does not depend on d, but on n. In a second example, we consider a matrix with positive spectrum included in $[0, 20]$. We use the shift method presented in Remark 3.3 to compute the exponential. We see that the error still does not depend on the dimension of the problem.

From now on, we focus on the matrix obtained by the usual second order Finite Difference discretization of the one dimensional Laplace operator, that we denote by $A = \Delta_d^1 \in \mathcal{M}_d(\mathbb{R})$. The relative error is computed in practise by $\frac{\|\texttt{expm}(\Delta_d^1)-\mathcal{R}_n(\Delta_d^1)\|_2}{\|\texttt{expm}(\Delta_d^1)\|_2}$. The results are presented in Figure 9 (left). Here again, the error does not depend on the dimension of the problem, but only on the degree of truncation n.

Next, we compare the computing time (using `tic` and `toc` functions of Matlab and Octave) of $\texttt{expm}(\Delta_d^1)$ and $\mathcal{R}_n(\Delta_d^1)$ denoted respectively t_{seq} and t_{para}. The results are presented in Figure 9 (right). In this test, the matrices $(\Delta_d^1 + \theta_k^{(n)} I)^{-1}$ are computed in parallel and t_{para} is defined as the maximum time taken to compute one of the $a_k^{(n)}(\Delta_d^1 + \theta_k^{(n)} I)^{-1}$. We can see that t_{seq} is slightly larger than t_{para} in the case of

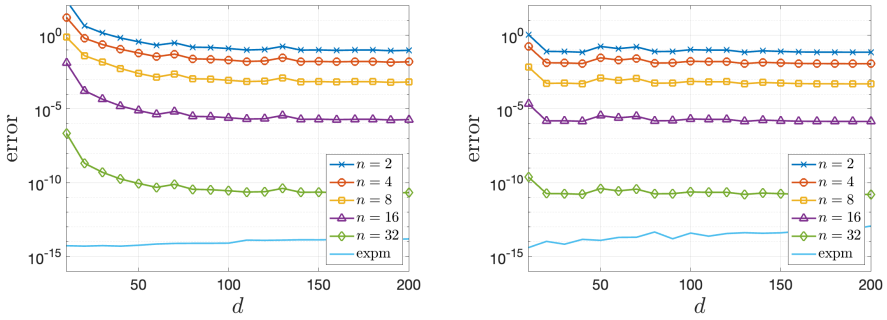

Figure 8: Relative error versus the dimension of the matrix. Left: matrices with negative spectra. Right: matrices with positive spectra using the shift method from Remark 3.3.

Matlab and almost ten times larger in the case of Octave.

We finally consider the action of the matrix exponential on vectors. For $v \in \mathbb{R}^d$, the vector $w = \exp(\Delta_d^1)v$ is approximated by (3.2) where $(\Delta_d^1 + \theta_k^{(n)} I)y_k^{(n)} = v$ is solved using the solvers `mldivide` of Matlab and Octave. We evaluate the mean of the error and the mean of t_{para} for a series of random vectors v.
The relative error $\frac{\|\texttt{expm}(\Delta_d^1)v - \mathcal{R}_n(\Delta_d^1)v\|_2}{\|\texttt{expm}(\Delta_d^1)v\|_2}$ together with the computing times t_{seq} and t_{para} for $\texttt{expm}(\Delta_d^1)v$ and for $\mathcal{R}_n(\Delta_d^1)v$, respectively are shown in Figure 10. In this case t_{para} is defined as the maximum time used to compute one of the vectors $a_k^{(n)} y_k^{(n)}$. We note that t_{seq} is larger than t_{para} for all values of the dimension d of the matrix, with Matlab as well as Octave. For example, with $d = 10^3$, $t_{seq} \approx 10 t_{para}$ and $t_{seq} \approx 10^5 t_{para}$ with Matlab and Octave, respectively.

6. Concluding remarks. We have presented a simple and efficient method to compute the exponential of a Hermitian matrix. This method is based on a rational approximation of the scalar exponential. We explain in particular how this idea, old and very well documented, is particularly suitable for parallel computing. The tests show a substantial computational time saving in matrix-vector products of the form $\exp(A)v$. The method also offers an advantage in terms of memory occupation. Indeed, in the general case, the matrix $\exp(A)$ is full and must be stored entirely for the direct calculation of the vector $\exp(A)v$. On the contrary, our method does not requires to store the full matrices $(A+\theta_k^{(n)} I)^{-1}$ but only the sparse matrices $A+\theta_k^{(n)} I$.

REFERENCES

[1] A. H. Al-Mohy and N. J. Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM Journal on Scientific Computing*, 33(2):488–511, 2011.

[2] G. Baker, P. Graves-Morris, and S. Baker. *Padé Approximants*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1996.

[3] D. Braess. On the conjecture of meinardus on rational approximation of $\exp(x)$. *J. Approx. Theory*, 36(4):317–320, 1982.

[4] W. Cody, G. Meinardus, and R. Varga. Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems. *Journal of Approximation Theory*, 2(1):50–65, 1969.

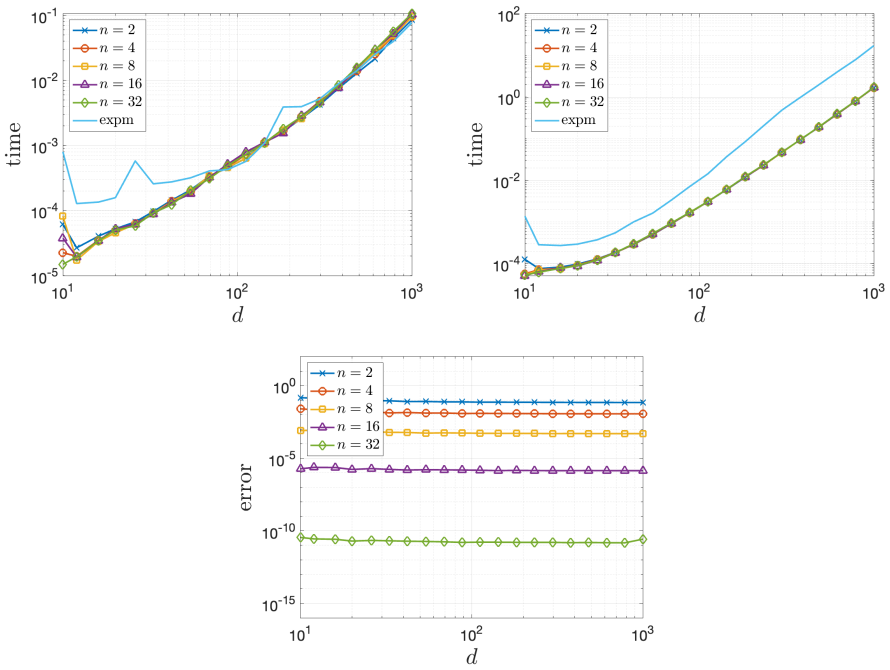

Figure 9: Same as Figure 8 for matrix Δ_d^1. Top, Left: corresponding CPU time required to compute `expm` and approximation $\mathcal{R}_n$ for various values of n using Matlab. Top, Right: same using Octave. Bottom: relative error in the computation of $\exp(\Delta_d^1)v$ for vectors v randomly chosen as a function of the dimension of matrix Δ_d^1.

[5] J. W. Eaton, D. Bateman, S. Hauberg, and R. Wehbring. *GNU Octave version 5.2.0 manual: a high-level interactive language for numerical computations*, 2020.

[6] E. Gallopoulos and Y. Saad. Efficient parallel solution of parabolic equations: Implicit methods on the cedar multicluster. In J. Dongarra, P. Messina, D. C. Sorensen, and R. G. Voigt, editors, *Proc. of the Fourth SIAM Conf. Parallel Processing for Scientific Computing*, pages 251–256. SIAM, 1989.

[7] Y. Y. Lu. Exponentials of symmetric matrices through tridiagonal reductions. *Linear Algebra and its Applications*, 279(1):317–324, 1998.

[8] The Mathworks, Inc., Natick, Massachusetts. *MATLAB version 9.11.0.1769968 (R2021b)*, 2021.

[9] G. Meinardus. *Approximation of Functions: Theory and Numerical Methods*. Springer tracts in natural philosophy. Springer, 1967.

[10] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[11] E. B. Saff and R. S. Varga. Zero-free parabolic regions for sequences of polynomials. *SIAM Journal on Mathematical Analysis*, 7(3):344–357, 1976.

[12] A. Schönhage. Zur rationalen approximierbarkeit von e^{-x} über $[0, +\infty)$. *Journal of Approximation Theory*, 7(4):395–398, 1973.

[13] G. Szegö. Über einige eigenschaften der exponentialreihe. *Sitzungsber. Berl. Math. Ges*, 23:50–64, 1924.

[14] L. N. Trefethen. The asymptotic accuracy of rational best approximations to e^z on a disk. *Journal of Approximation Theory*, 40(4):380–383, 1984.

[15] S. M. Zemyan. On the zeroes of the $n-th$ partial sum of the exponential series. *The American Mathematical Monthly*, 112(10):891–909, 2005.

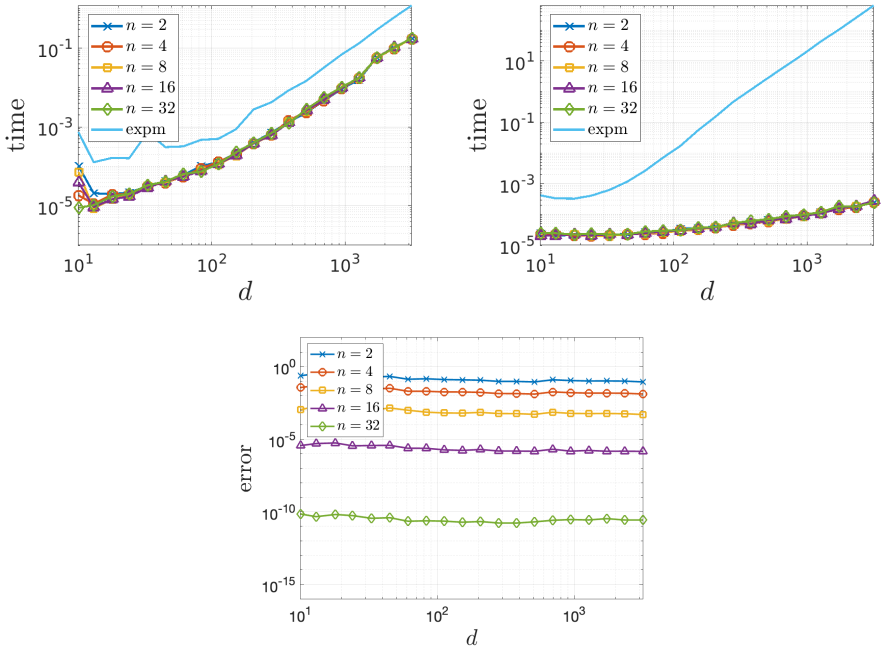

Figure 10: Top, Left: corresponding CPU time required to compute `exmp` and approximation $\mathcal{R}_n$ for various values of n using Matlab. Top, Right: same using Octave. Bottom: relative error in the computation of $\exp(\Delta_d^1)v$ for vectors v randomly chosen as a function of the dimension of matrix Δ_d^1.